III Statistics

Statistics has become the essence of many modern applications, especially in the tech world and in finance. There was never a time in history where we were able to obtain as much data as we currently do, and therefore a deep knowledge of statistics has become the most demanded qualification in today's job market. The tools of collecting, analyzing, and finding nontrivial relations between data is what gives some very successful companies their edge. For example, in the financial industry, there are a select few hedge funds that always beat the market. The reason this happens is because those funds have found relations between data that the rest of the market hasn't, which in turn allow them to make profits that the rest of the market can't. In the sections that follow we touch upon a few topics from the vast field of statistics.

III.1 Data and Distributions

There are two types of variables: numerical and categorical. Numerical variables can be classified as either discrete or continuous. For example, the number of books you own is a discrete numerical variable, your height, on the other hand, is continuous. Categorical variables can also be divided into two categories: either regular or ordinal. For example, when tossing a coin, the outcomes *Heads* and *Tails* are considered regular categorical variables, but when responding to a survey that asks you to rate a certain service by selecting whether you were *very satisfied, somewhat satisfied, somewhat dissatisfied,* or *very dissatisfied*, then these are considered ordinal categorical variables. A dataset usually consists of a collection of numerical and/or categorical variables.

Looking at raw data in a dataset and trying to draw meaningful information from their values will most likely prove futile. Therefore, one of the most important tools in analyzing data is having a good visualization technique. There are several ways to visualize data, and here we will only discuss two: scatter plots and histograms.

- Scatter plots are two dimensional plots that show the relation between two variables. An example of a scatter plot can be seen in Figure 1, which shows the relation between the life expectancy of individuals in 180 countries and the average income per person (GDP per capita) in that country (the dataset was obtained from gapminder.com). This plot clearly shows that people who make more money tend to live longer. Unfortunately, this function levels off at around 85 years, because humans haven't found a way to live forever (yet)!
- A histogram shows all the possible data outcomes of a given variable and how often they occur. For example, we can use the above dataset to see whether more countries

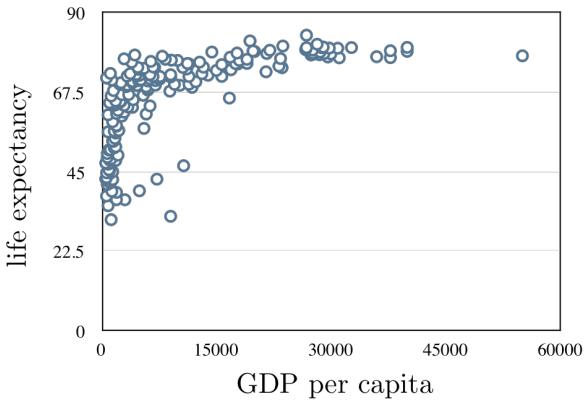


Figure 1: Scatter plot of life expectancy as a function of income per capita.

have high or low life expectancy. The histogram in Figure 2 shows that the majority of countries have a life expectancy above 60 years, with the highest frequencies being between 70 and 80 years. Since more people live longer, the histogram is left-skewed, i.e. it has a long tail on the left.

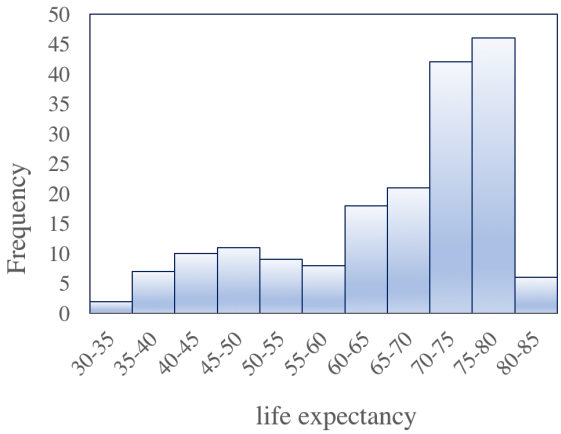


Figure 2: Histogram of life expectancy from 180 countries.

III.2 Important parameters of distributions

There are several parameters through which a distribution can be characterized. We review a few of those below:

mean: the mean, or arithmetic average, of a set of observations $x_1, \ldots, x_n$ is given by

$$\overline{x} = \frac{x_1 + \cdots + x_n}{n}. \tag{1}$$

median: the median of a set of observations $x_1, \ldots, x_n$ is calculated by first arranging the set in ascending order, and then selecting the midpoint or center value if the number of elements n is odd or the average of the two center values if n is even.

sample variance: The variance is the average of the square deviations of the data from the mean. If we have a set of observations $x_1, \ldots, x_n$, with a mean $\overline{x}$, then the variance is given by

$$s^2 = \frac{(x_1 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1}. \tag{2}$$

sample standard deviation: The standard deviation is the square root of the variance, and it could be thought of as an error on the mean.

modality: A distribution could be uniform (no peaks), unimodal (one peak), bimodal (two peaks), or multimodal (many peaks) as shown in Figure 3.

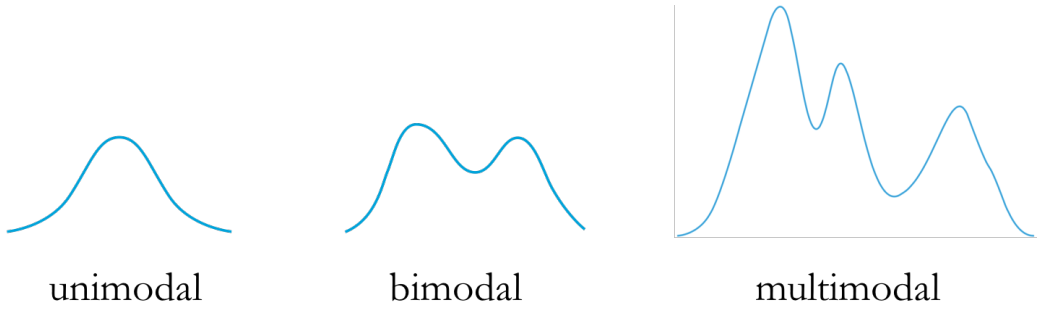


Figure 3: Varying modality.

skew: A distribution could be left-skewed, right-skewed, or symmetric (no skew). For a left-skewed distribution, the mean is typically smaller than the median; for a right-skewed distribution, the mean is larger than the median; and for a symmetric distribution the mean is approximately equal to the median. These results are summarized in Figure 4.

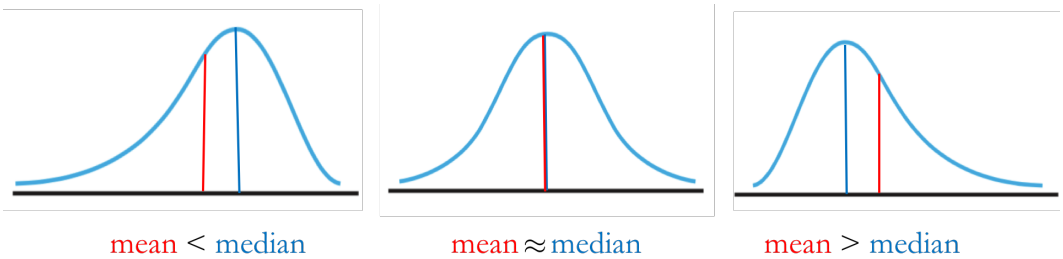


Figure 4: How skew affects the relation between the mean and the median.

Example: A wildlife biologist measures the lengths (in centimeters) of adult male and female stoats captured in England:

male stoats:	20, 21, 23, 16, 20, 23, 25, 19, 18, 18, 20, 21, 22, 24, 18, 19, 20
sorted:	16, 18, 18, 18, 19, 19, 20, 20, 20, 20, 21, 21, 22, 23, 23, 24, 25
female stoats:	14, 20, 23, 19, 13, 22, 22, 20, 19, 13, 18, 17, 15, 20
sorted:	13, 13, 14, 15, 17, 18, 19, 19, 20, 20, 20, 22, 22, 23

	M	F
Mean	20.4	18.2
Sample Standard Deviation	2.3	3.2
median	20	$\frac{19+19}{2} = 19$

III.3 The Normal Distribution

The most important distribution in statistics is called the normal or Gaussian distribution. What makes it very special are the rules that this distribution obeys which make many calculations very easy to perform. There are no known variables that have an exact normal distribution, but many numerical data can be approximated by the normal curve. Examples include heights, weights, and IQ tests. Even when the variables are distributed very far from normal, their averages will be normally distributed under the right conditions. This will be discussed in more detail in the section on the Central Limit Theorem.

The normal distribution is shaped like a bell, it has a mean μ, a standard deviation σ, and is typically denoted by $N(\mu, \sigma)$. The distribution is symmetric about μ, and it follows a few very strict rules about how the data is distributed: 68% of the data is within one standard deviation of the mean, i.e. 68% of the data lies between $\mu - \sigma$ and $\mu + \sigma$; 95% is within 2 standard deviations, i.e. 95% of the data lies between $\mu - 2\sigma$ and $\mu + 2\sigma$; and 99.7% of data is within 3 standard deviations of the mean, i.e. 99.7% of the data lies between $\mu - 3\sigma$ and $\mu + 3\sigma$. This is referred to as the 68-95-99.7 rule, and is summarized in Figure 5.

The standard normal distribution

The standard normal distribution, also called the z-curve, has a mean $\mu = 0$, standard deviation $\sigma = 1$, and is denoted by $N(0, 1)$. A variable that has a standard normal distribution is called a **standard normal variable** and is typically denoted by Z.

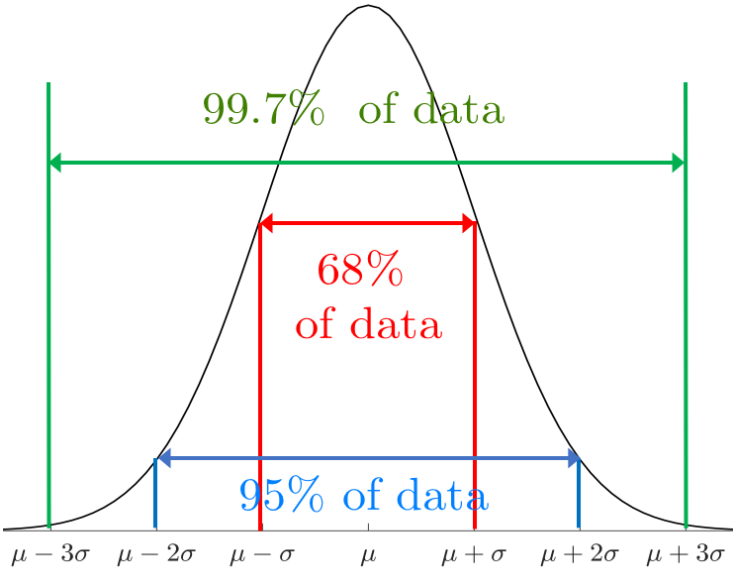


Figure 5: The 68–95–99.7 rule for normal distributions.

A variable X having a normal distribution $N(\mu, \sigma)$ can be **standardized** by performing a change of variables that transforms X into a standard normal variable Z. This transformation is usually very useful when calculating probabilities under a normal distribution using the standard normal table (attached at the end of this document), or when calculating percentiles. To standardize $X \sim N(\mu, \sigma)$, we calculate

$$Z = \frac{X - \mu}{\sigma}, \tag{3}$$

where Z is a standard normal variable called the z-score. Although this idea stems from standardizing a non-standard normal variable, we can actually calculate the z-score for any distribution with mean μ and standard deviation σ. In general, a z-score is defined as

$$Z = \frac{\text{observation} - \mu}{\sigma}. \tag{4}$$

Calculating probabilities and percentiles

Suppose you have a random variable X which is normally distributed, with $X \sim N(\mu, \sigma)$, and suppose you are interested in finding the probability that X is smaller than some number x, i.e. $P(X \leq x)$. This probability corresponds to the area under the normal curve to the left of x, and it can be calculated by first standardizing X and turning it into $Z = (X - \mu)/\sigma$, which turns x into $z = (x - \mu)/\sigma$. Then the probability $P(X \leq x)$ becomes $P(Z \leq z)$, which is shown in Figure 6. The calculation can be performed using the standard normal table (attached at the end of this document) which has z-scores up to two decimal points. To read

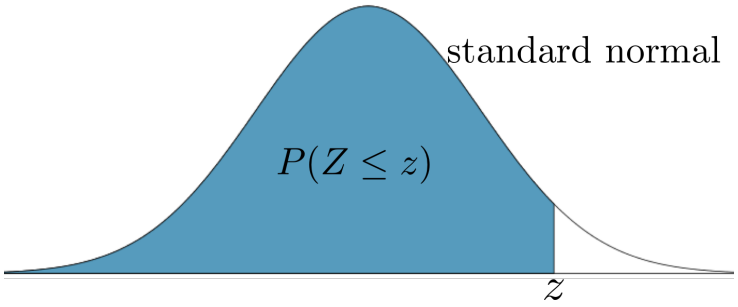


Figure 6: A standard normal distribution where the shaded area corresponds to $P(Z \leq z)$.

the table for a given z-score, we find the first decimal point from the left-most column of the table, then we find the second decimal point from the top column. For example, if we want to calculate $P(Z \leq 2.81)$, we must first locate "2.8" in the left-most column, then from the top we find ".01", and we determine where the "2.8" row and the ".01" column intersect, which turns out to be 0.9975. Then,

$$P(Z \leq 2.81) = 0.9975. \tag{5}$$

A percentile of the standard normal distribution corresponds a value on the horizontal axis, such that the area under the z-curve to the left of that value equals the percentile value divided by 100. For example, the 95th percentile corresponds to a value z, such that the area under the curve to the left of z equals 0.95. Similarly, the 5th percentile corresponds to a value with an area of 0.05 to the left of it. This is the inverse of the calculation $P(Z \leq z) =?$ performed above. Now we have the right-hand-side, and we need to find z, so we have to use the table in reverse. For example, if you want to calculate the 99th percentile, you would write

$$P(Z \leq z) = 0.99, \tag{6}$$

and then you would look for the closest number to 0.99 in the body of the table, and read off its z-score. The closest number is 0.9901 and it can be found in the second table in the 4th column and 24th row, which corresponds to a z-score of 2.33.

Example: A student scores 78 on an exam which was approximately normally distributed according to $N(68, 5)$.

- What is the percentile score of this student?

$$P(X \leq 78) = P\left(\frac{X - 68}{5} \leq \frac{78 - 68}{5}\right) = P(Z \leq 2) = 0.9772 \approx 0.98. \tag{7}$$

Therefore, this student scored in the 98th percentile.

- Which grade corresponds to the 99th percentile?

$$P(Z \leq z) = 0.99 \implies z = 2.33. \tag{8}$$

But $z = (x - \mu)/\sigma$, and therefore $x = (\sigma \times z) + \mu = (5 \times 2.33) + 68 = 79.65$.

The Central Limit Theorem

Suppose you have a population with any given underlying distribution, which has a mean μ and standard deviation σ, and suppose that you take n samples $\{X_1, X_2, \ldots X_n\}$ from this distribution, and you calculate the average value of each sample $\{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n\}$, where $\overline{x}_1$ is the average value of the sample X_1, etc.. Then, the distribution of the averages $\{\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_n\}$, which we call $\bar{X}$, will be normal with mean μ and standard deviation $\sigma/\sqrt{n}$ if the following conditions are met:

1. The sampled observations must be independent.
2. If sampling without replacement, $n < 10\%$ of population.
3. If the underlying population distribution is skewed, the sample size can be small. If it is non-normal, the sample size must be large (rule of thumb: $n > 30$).

The above is a statement of the **Central Limit Theorem** which we illustrate further with an example.

Example [This problem is selected from Chapter 4 in: Devore, J. L., *Probability and Statistics*, 9th Edition (2016)]
Consider the distribution shown in Figure 7 for the amount purchased (rounded to the nearest dollar) by a randomly selected customer at a particular gas station. The distribution is obviously quite non-normal.

We asked a computer program to select 1000 different samples, each consisting of $n = 15$ observations, and calculate the value of the sample mean $\bar{X}$ for each one. Figure 8 is a histogram of the resulting 1000 values; this is the approximate sampling distribution of $\bar{X}$ under the specified circumstances. This distribution is clearly approximately normal even though the sample size is actually much smaller than 30, our rule-of-thumb cutoff for invoking the Central Limit Theorem. It is typically not non-normality of the population distribution that causes the CLT to fail, but instead very substantial skewness.

One final remark about the central limit theorem: note that the distribution of the sample mean has a standard deviation equal to $\sigma/\sqrt{n}$, therefore, the larger the sample size, the

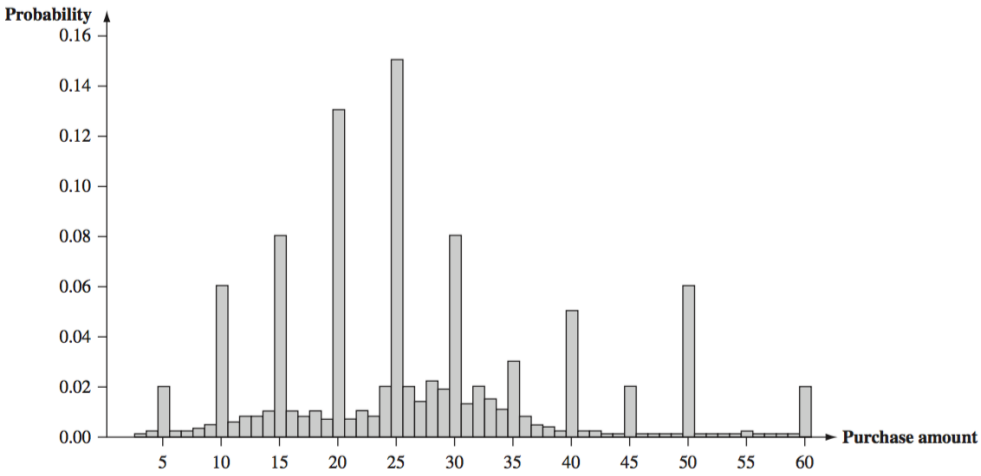


Figure 7: Probability distribution of X amount of gasoline purchased ($).

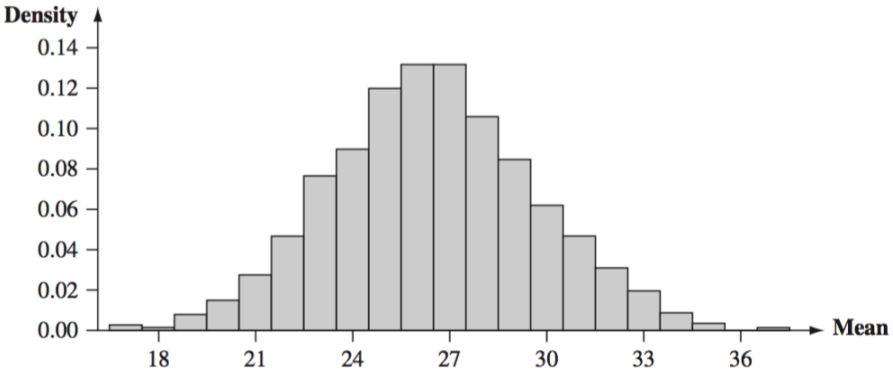
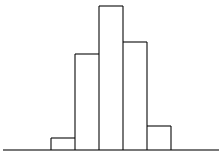


Figure 8: Approximate sampling distribution of the sample mean amount purchased when $n = 15$.

smaller the standard deviation or spread, and the more concentrated the data will be around the mean μ.

Remarks about normal distributions

In practice, many observations can be approximated using the normal distribution. For example, if you try to measure a quantity really carefully, say the mass of an electron, then you can't trust one single measurement as being accurate: your measurement will also contain some unavoidable error (because, however carefully you calibrate all your apparatus, they will never be "infinitely" precise). So then you repeat the measurement several times, and you plot the observed values. Typically, they cluster around some central value, and when you make a histogram around this central value, indicating how many of the observations fell within a central bin, or in the next bin, or even further, ... you get something that looks close to a normal distribution, such as:



How can we explain this? After all, we are not averaging anything here? The error itself, however, is a conglomerate of many different things (none of which we can control, otherwise we would get rid of it!—With technological advances, we can and do indeed reduce the error when we repeat those classical measurements). The total error can therefore be viewed as the sum total of all these different influences. On this sum the central limit theorem plays its role → normal distribution.

What about for other types of measurements? For instance, the height of young men between 25 and 30 in the US? Or the number of calories in people's diets? Or how well they see small printed characters?

Here again, in a population where there are no obvious inhomogeneities (and here this is a big assumption that will need to be carefully examined), one usually observes things which look like normal distributions around a central value. It is believed that the deviation from the average is caused by many different independent factors. (Do you really believe this? Isn't height, for instance, largely determined by genetics? How would you then still explain observing a normal distribution?)

Simpson's paradox

Simpson's paradox illustrates how easily one can be misled by statistical consideration of scenarios in which one has an aggregate of individuals in a variety of different circumstances.

In the 105/104 academic school year of the University of Alexandria, the Thracians got angry at the Admissions Office, because they found out (and it was an incontrovertible fact) that 4400 Dacians had applied and 4400 Thracians had also applied, but 3280 Dacians were admitted while only 1120 Thracians were admitted.

We can present this data on a table:

Nationality	Applicants	Admitted	Percent Admitted
Dacian	4400	3280	74.55 %
Thracian	4400	1120	25.45 %

However, it was pointed out to the Thracians that there were two academic divisions in U. of Alexandria, the Trivium (Grammar, Logic, and Rhetoric), and the Quadrivium (Arithmetic, Geometry, Astronomy, and Music). If we break down the numbers according to division, then we see a different picture emerge:

Nationality	Triv. Applicants	Triv. Acceptances	Triv. % Admitted
Dacians	4000	3200	80 %
Thracians	400	320	80 %
Nationality	Quadriv. Applicants	Quadriv. Acceptances	Quadriv. % Admitted
Dacians	400	80	20 %
Thracians	4000	800	20 %

Check that if you total up the number of Dacian applicants to both divisions, you get a total of 4400 and if you total up the number of Dacians admitted, you get 3280. Likewise, the totals for Thracians are exactly as we originally said. Yet, now it would appear that there is no bias at all in the admissions process—in each division the admission rates are the same for Dacians and Thracians, and more Dacians get in because they flock to the easier division (Trivium). This is an example of a "hidden variable," namely, that different divisions have different admission rates, and that Thracians seem to prefer the harder division.

Then a clever Dacian realized that there might be further hidden variables. He realized that the Trivium has two departments, one handling Grammar and Logic (G&L) and the other handling Rhetoric. And likewise, the Quadrivium is split into two departments: Arithmetic, Geometry, and Astronomy (affectionately known as AGA), and the Department of Music. The Dacian split the data further.

Trivium Programs

Nationality	G&L Applicants	G&L Acceptances	G&L % Admitted
Dacians	3600	3060	85 %
Thracians	100	95	95 %
Nationality	Rhetoric Applicants	Rhetoric Acceptances	Rhetoric % Admitted
Dacians	400	140	35 %
Thracians	300	225	75 %

Quadrivium Programs

Nationality	AGA Applicants	AGA Acceptances	AGA % Admitted
Dacians	300	75	25 %
Thracians	400	260	65 %

Nationality	Music Applicants	Music Acceptances	Music % Admitted
Dacians	100	5	5 %
Thracians	3600	540	15 %

Again, one can check that the total number of Dacians who applied to Trivium programs (G&L and Rhetoric) is 4000, of whom 3200 were admitted. This agrees with the last analysis. And likewise you can check that all the other totals for Dacians or Thracians in Trivium or Quadrivium are as in the previous analysis. But the further breakdown shows that Thracians had a higher admission rate in EVERY SINGLE department, despite the fact that the overall admission rate for Thracians was much lower than that of the Dacians!

So the initial Thracian uproar seems unjustified, to say the least. Until we find another hidden variable...

When do we know that we have found all the hidden variables? In general, we don't. In any case, finding hidden variables here was not done by a mathematical procedure, but rather by reflection on factors present in the situation being analyzed (in this case, the structure of the University of Alexandria).

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0017	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0038
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
−1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
−1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
−1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
−1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
−0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
−0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
−0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
−0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
−0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
−0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
−0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3482
−0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
−0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
−0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9278	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998